BAN 620-01 DATA MINING PROJECT PROPOSAL

Comprehensive Data Mining Approach to Predict Startup Success and Revenue Using Machine Learning Techniques

Project Description:
This project aims to apply a complete suite of data mining methods to analyze and model startup performance using a real-world dataset of 5,000 startups. The dataset includes numerical information such as funding amount, revenue, employee count, burn rate, and customer retention. With this data, we aim to tackle both predictive and exploratory tasks:

1. Predict startup success/failure using classification models.

2. Predict expected revenue using regression models.

3. Segment startups using clustering techniques.

4. Discover risk and growth patterns using association rule mining.

5. Simulate recommendation scenarios for startup-investor pairing via collaborative filtering.

Data Mining Techniques to be applied:

● Multiple Linear Regression: Predict startup revenue using numerical predictors like burn rate and marketing expenses.
● k-Nearest Neighbors (k-NN): Classify startup success or predict revenue based on similar startups.
● Classification & Regression Trees / Ensemble Trees (e.g., Random Forest): Identify interpretable business rules and improve prediction accuracy.
● Logistic Regression: Model the probability of startup failure.

● Neural Networks: Learn complex patterns in startup data for classification and regression.

● Cluster Analysis (e.g., K-Means): Segment startups based on growth metrics, market size, and product uniqueness.

- Collaborative Filtering: Simulate recommendation of startups to investors based on potential fit, using startup feature similarity.

Link to the Dataset:

https://www.kaggle.com/datasets/sakharebharat/startup-failure-prediction-dataset

Tools & Libraries:

Python, with packages: scikit-learn, pandas, matplotlib, seaborn, and more.

Group A includes:

Anantha Vyasa Kurudi (Wf2471)
Gopi Krishna Reddy Katkuri (Rn7945)
P. Pooja Prasad (jl6843)
Sanjana Paluru (On1935)
Shraddha Sudhir (pb8129)